BhiliVaani: Dialect-Faithful TTS for a Low-Resource Indian Tribal Language via Staged Cross-Lingual Augmentation

Anonymous authors
Paper under double-blind review

Abstract

Endangered languages often lack text-to-speech (TTS) systems due to scarce paired speech data. We study Bhili, an Indian tribal language without existing TTS resources, and introduce **BhiliVaani**, consisting of Core (community utterances) and Bridge (synthetic utterances) corpus. Bridge audio is highly intelligible ($\text{MOS}_{\text{Int}} = 4.86 \pm 0.08$) but exhibits Hindi-accented prosody, motivating a two-axis benchmark evaluating intelligibility and dialect authenticity. We benchmark VITS, GlowTTS, and AlignTTS across five regimes: Original, Synthetic, Mixed, Syn→Org, and staged Org→Syn with incremental synthetic scaling (+2,+4,+6 h). Evaluation includes community $\text{MOS}_{\text{Int}}/\text{MOS}_{\text{Dial}}$, MUSHRA, diagnostic metrics (VAD-SNR, ΔC50, F0), and manual error analysis. Results show that increasing synthetic data improves clarity but can reduce dialect authenticity. The Org→Syn staged strategy reveals a controllable trade-off frontier and a preservation-oriented approach balancing intelligibility and dialect fidelity.

1 Introduction

The Indian subcontinent is among the world's most linguistically diverse regions, with 1,369 rationalized mother tongues and 22 constitutionally recognized scheduled languages, as reported in the Census of India 2011[1] . Despite this diversity, the speech technology ecosystem remains concentrated around a small set of high-resource languages. Mainstream commercial TTS and ASR services support only dozens of languages, leaving the majority of India's low-resource and tribal languages without robust speech-enabled interfaces. This gap is consequential because speech technologies increasingly mediate access to public information and services, education, telephony and Interactive Voice Response (IVR)systems, and accessibility tools for visually impaired users Kim et al. (2024) Ruder et al. (2019). The mismatch between linguistic coverage and technological support is particularly acute for tribal languages. India has approximately 443 tribal languages[2], many of which remain under-documented and digitally sparse. UNESCO reports that 30 Indian languages are critically endangered[3], and endangerment is often driven by *domain shift*: younger speakers adopt digitally supported languages for schooling, employment, and online communication, while the heritage language contracts to private or ceremonial contexts. In such settings, language revitalization is tightly coupled with digital inclusion: if a language is absent from speech technologies, it is effectively absent from the interfaces through which contemporary services are delivered Srinivasa Varadhan et al. (2024). Community discussions further emphasize that language mismatch can discourage engagement with welfare initiatives and digital governance, and can contribute to educational disengagement when learning materials and platforms do not reflect local language and cultural contexts Kumar et al. (2023).

1.1 Bhili: India's Largest Tribal Language without TTS Infrastructure

We study Bhili, among the largest tribal languages in India, spoken by approximately 17 million people across Rajasthan, Madhya Pradesh, Gujarat, and Maharashtra[4]. Despite its scale and cultural significance,

[1] https://censusindia.gov.in/nada/index.php/catalog/10191
[2] https://www.pib.gov.in/PressReleseDetailm.aspx?PRID=1656686&reg=3&lang=2
[3] https://www.unesco.org/en/articles/towards-world-atlas-languages
[4] https://en.wikipedia.org/wiki/Bhil

Bhili has remained largely invisible in modern speech technology, publicly available speech-text resources are scarce and there is no widely adopted open benchmark for Bhili TTS. We focus on the Jhabua dialect written in Devanagari and adopt a community-centered framing where the *oral* dimension is central, that means pronunciation, rhythm, prosody, and performance traditions (e.g., storytelling and song) are difficult to preserve through text alone. For first-generation learners, low-literacy speakers, visually impaired users, and multilingual classrooms, reliable audio is therefore foundational Lux et al. (2022).

Text-to-speech (TTS) offers a distinctive pathway to scale in this setting, once a voice model exists, it can generate unbounded speech from text, enabling pronunciation-aware learning materials, audio-first public information, and voice-based interfaces Kumar et al. (2023). However, neural TTS typically requires tens to hundreds of hours of paired speech-text data Sankar et al. (2024), which is difficult to collect for tribal languages due to limited studio access and the opportunity cost of large-scale recording. A common response is to leverage related-language resources or synthetic augmentation Pine et al. (2022). Yet in revitalization-oriented scenarios, quality is not only intelligibility: synthetic speech can import non-local accent and prosody, yielding outputs that are clear but socially *misrecognized* as sounding like a dominant neighboring language. For Bhili, Devanagari overlap with Hindi makes related-language bootstrapping feasible but also increases the risk of Hindi-like prosodic leakage and dialect dilution.

This motivates our central question: *How can we build a useful, dialect-aware Bhili TTS system under extreme data scarcity, using augmentation to improve intelligibility without diluting dialect identity?* We hypothesize that when the available real speech is insufficient to train an acceptable system, controlled synthetic augmentation can complement the limited real recordings. In this setup, the real Bhili data primarily provides dialectal cues, while synthetic speech carefully mixed during training targets improved clarity and intelligibility.

We introduce **BhiliVaani**, a community-driven speech resource and benchmark for Bhili, and an empirical study of training strategies that explicitly regulate the intelligibility-dialect trade-off. Concretely, our contributions are:

- We curate a high-quality Bhili speech-text dataset comprising 3 hours of original Bhili audio recordings and 6 hours of synthetic speech generated from Bhili text using a Hindi-TTS pipeline, producing a total of 9 hours of training data from 3 speakers. This is, to our knowledge, the first open TTS-oriented dataset of this scale for Bhili.
- We introduce a Bhili TTS benchmark split and an evaluation framework that separates intelligibility from dialect authenticity, reflecting revitalization-oriented requirements.
- We systematically compare multiple training regimes for non-autoregressive TTS models (VITS, GlowTTS, AlignTTS), including original-only, synthetic-only, mixed-data training, synthetic-to-original fine-tuning, and a proposed staged dialectal refinement curriculum.
- Empirical findings with community-facing relevance: Using MOS-based listening tests and targeted analysis of synthesis errors, we show that strong performance is achievable with limited real bhili speech data when synthetic augmentation is introduced carefully. In our experiments, the best systems achieve MOS around 4.20 (intelligibility) and 3.60 (dialect), indicating practical usability in downstream applications.

1.2 Speech Technology for Language Revitalization

For many Indigenous and tribal communities, oral transmission remains central: speech carries pronunciation, prosody, rhythm, and performance traditions that are difficult to preserve through text alone Keith (2024). While documentation initiatives are expanding, community stakeholders repeatedly emphasize that the *audio-first* dimension of preservation remains underserved Wang et al. (2025). For Indian tribal languages, this gap has consequences beyond cultural heritage. When speech interfaces do not exist in a community's language, accessing public information and services often requires navigating text-heavy workflows in dominant languages Casanova et al. (2022); similarly, learners may disengage when educational tools do not reflect local language realities. TTS is well-suited to address these constraints because it converts limited

textual resources into scalable audio, enabling pronunciation-aware learning materials, audiobooks, classroom prompts, accessible interfaces, and voice-based public information systems Chen et al. (2019). Importantly, TTS should be viewed as *infrastructure rather than replacement*: community-aligned voice models can amplify carefully recorded native speech while reducing the recording burden on elders and educators. Critically, "quality" in revitalization-oriented TTS is not equivalent to intelligibility alone Bird (2020). If transferred or synthetic speech becomes locally *misrecognized* as sounding like a dominant neighboring language, adoption can be undermined even when the content is understandable. This motivates evaluation that explicitly separates intelligibility from perceived dialect authenticity Boito et al. (2018). BhiliVaani provides a controlled empirical case study of a common low-resource synthesis setting: authentic target-language speech is scarce, while synthetic speech from a related high-resource language is available. In this setting, synthetic supervision can improve intelligibility but may also transfer donor-language suprasegmental patterns, thereby reducing perceived dialect authenticity. We therefore treat training order and synthetic-data exposure as controlled experimental variables for analysing dialect drift, rather than as merely implementation-level design choices.

2 Related Work & Background

Neural text-to-speech systems require paired speech-text supervision to learn stable alignment, pronunciation, and prosody. Autoregressive systems such as Tacotron-style Elias et al. (2021) models can produce natural speech in high-resource settings, but their attention mechanisms are fragile when only a few hours of target-language audio are available, often causing skipped, repeated, or unintelligible output. Non-autoregressive architectures such as FastSpeech, Glow-TTS, AlignTTS, and VITS reduce this instability through duration modeling, monotonic alignment search, or end-to-end acoustic generation, and are therefore widely used in low-resource TTS studies Ueda et al. (2024). Nevertheless, architecture alone does not remove the data bottleneck: with extremely limited native speech, systems may preserve some local acoustic cues but fail to produce consistently intelligible synthesis.

A major line of work addresses this bottleneck through multilingual training Casanova et al. (2024), cross-lingual transfer Badlani et al. (2022), phone mapping, source-language selection, and synthetic augmentation Bhushan et al. (2025). These methods are attractive when a related high-resource language shares script or phonological structure with the target languageConneau et al. (2020). However, cross-lingual supervision is not neutral. Donor-language systems can transfer not only segmental information but also suprasegmental priors such as rhythm, intonation, and phrasing. As a result, synthetic speech may improve intelligibility while making the output sound less locally authentic. This trade-off is central to our setting: Hindi synthesis can render Bhili text largely understandable, but may also introduce Hindi-accented prosody. We therefore treat synthetic speech as regulated supervision whose amount and training order must be explicitly evaluated.

Indian-language TTS has advanced through systems and resources such as Indic-TTS Kumar et al. (2023), IndicVoices-R Sankar et al. (2024), Indic Parler-TTS Sankar et al. (2025), Rasa Srinivasa Varadhan et al. (2024), LIMMITS Singh et al. (2024), and ELAICHI Anand et al. (2024), which improve multilingual coverage, speaker diversity, expressive synthesis, voice cloning, and rare-character handling for scheduled or comparatively resourced languages Pratap et al. (2024). Yet tribal and unscheduled languages remain largely outside this ecosystem. For Bhili, existing resources are primarily textual and can support prompt selection, but they do not provide the paired acoustic supervision required for neural TTS. BhiliVaani addresses this gap by combining a small dialect-grounded Bhili speech corpus with Hindi-derived synthetic speech, and by evaluating both intelligibility and perceived dialect authenticity Zhang et al. (2025). This positions our work at the intersection of low-resource TTS, cross-lingual synthetic augmentation, and preservation-oriented speech technology.

3 Data Construction

Neural TTS systems typically require tens to hundreds of hours of paired speech–text data for stable alignment and high perceptual quality Bird (2020). For Bhili, no widely used public TTS benchmark existed and

curated parallel speech–text resources were extremely limited. We therefore construct two complementary resources that separate (i) a *dialect-grounded* community corpus and (ii) a *controlled* synthetic corpus that increases training scale.

Field Constraints for Tribal Language Data Collection. Speech data collection for tribal languages in India presents challenges beyond typical low-resource settings. Many communities are geographically dispersed with limited infrastructure; dialect variation is substantial; and sociolinguistic factors such as dominant-language schooling reduce literacy and promote code-switching. For Bhili, dialectal heterogeneity across regions makes it essential to collect a coherent dialect variety rather than an inadvertent mixture that could confound model learning. Identifying speakers with consistent dialect-faithful usage and producing stable read speech requires sustained community engagement and linguistic vetting. Logistical constraints (portable equipment, non-studio acoustics, and intermittent electricity) further increase the cost of each hour of paired speech–text data, alongside the governance and consent processes required for acceptable downstream use.

Table 1: Parallel Hindi–Bhili Sentence Examples.

ID	Hindi	Bhili
1	इनके पास सभी सम्मान के अधिकार हैं	इनी कन बदा इज्जत ना अधिकार से
2	छोटे परिवार वाले वे देश भक्ति को ही प्रकट करते हैं	नाना परिवार वाळा त्यां देह भक्ति ने प्रकट करे से
3	उन्होंने छोटे परिवार को रखकर भी देश भक्ति को ही प्रकट किया	तिहुयीं नाना परिवार ने राखी ने बी देह भक्ति नो थी प्रकट करियो
4	वे राष्ट्रभक्ति का प्रदर्शन करते हैं	त्यां देह नी भक्ति ना हळमा करे से

Table 2: Lexical and Morphological Correspondences between Hindi and Bhili.

Hindi	Bhili	Change type
इनके	इनी	pronoun simplification
सभी	बदा	lexical replacement
का/के/को	ना/ने/नी	postposition shift
हैं	से	copula change
करते हैं	करे से	habitual construction
किया	करियो	past suffix
छोटे	नाना	adjective substitution
वे	त्यां	pronoun shift
देश	देह	phonological shift
भी	बी	vowel shortening

BhiliVaani-Core: Original Bhili Speech Recordings. Through fieldwork in collaboration with local language activists and community organizations, we assembled 3 hours of original Bhili speech from the Jhabua district. Recordings were captured at 22,050 Hz, 16-bit PCM WAV. Prompt text was drawn from the BHEPC corpus Singh et al. (2025): we initially selected 2,000 transcriptions, applied a maximum text length of 270 characters, and discarded utterances longer than 10 s or shorter than 2 s.

After filtering, BhiliVaani-Core contains 1,700 utterances, partitioned into 1,000 training, 200 validation, and 500 held-out test utterances. The validation and test partitions are fully disjoint: validation utterances are used only for checkpoint selection, while the held-out test set is reserved exclusively for final perceptual and objective evaluation. Three speakers (2 male, 1 female) were born and raised in Jhabua, acquired Bhili as L1, and reported daily use. Screening targeted dialectal consistency and minimal Hindi phonological transfer. Transcription and segmentation were performed by trained Bhili linguists and verified by multiple native speakers to ensure orthographic consistency and clean text-audio pairing. Although modest in scale, BhiliVaani-Core provides the dialect-grounded anchor used across all training regimes and constitutes, to our knowledge, the largest phonetically aligned Bhili speech corpus currently available Ahumada et al. (2022).

Table 3: Phoneme inventory overlap between Hindi and Bhili. Note that inventory sizes reflect observed phoneme tokens produced by the eSpeak grapheme-to-phoneme converter and therefore depend on corpus coverage and phonemization granularity rather than a fixed theoretical phoneme set. Consequently, we focus on relative overlap metrics for comparison.

Hindi inventory size	142
Bhili inventory size	162
Shared phonemes	135
Coverage (Bhili covered by Hindi)	83.3%
Jaccard similarity	0.80

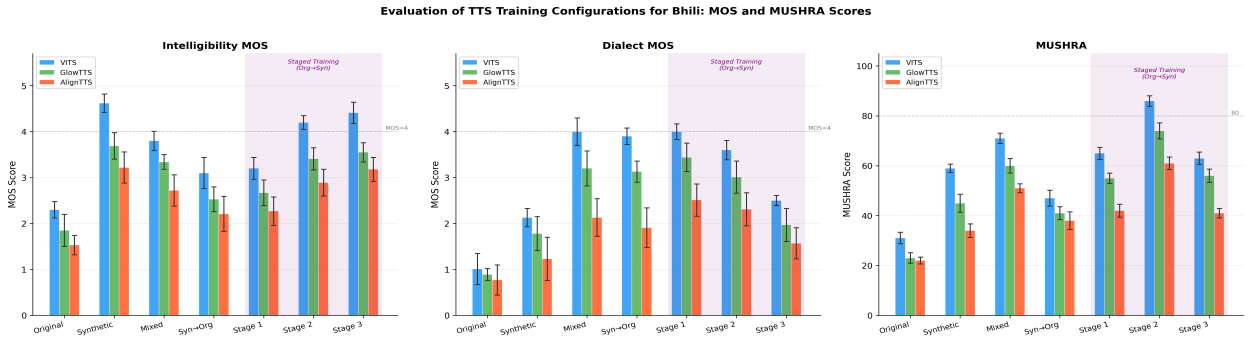

Figure 1: Intelligibility MOS, Dialect MOS, and MUSHRA scores across all training configurations for VITS, GlowTTS, and AlignTTS. The purple-shaded region highlights the Original→Synthetic staged training configurations. Error bars indicate standard deviation.

Scratch Training Reveals a Dialect-Intelligibility Trade-off. To assess feasibility under extreme scarcity, we trained neural TTS models from scratch using only BhiliVaani-Core. Across initial experiments with VITS, GlowTTS, and AlignTTS, outputs preserved recognizable Jhabua Bhili coloration and dialect cues, but intelligibility was limited: phonemes blurred, word boundaries were unstable, and prosody lacked sentence-level coherence. This reflects a core limitation of low-resource TTS: limited data may capture dialect identity, but often fails to provide sufficient supervision for robust alignment and clarity. This motivates the remainder of the paper: *how can we improve intelligibility without erasing dialect identity?*

Cross-Linguistic Augmentation: The Hindi Hypothesis. A practical strategy in data-scarce settings is to leverage related-language resources Baevski et al. (2020). Bhili and Hindi share Indo-Aryan ancestry and the Devanagari script, suggesting that Hindi-based synthesis may render Bhili text intelligibly even when suprasegmentals remain Hindi-like. We therefore explored synthetic augmentation via Hindi TTS as a mechanism to increase training volume.

Hindi-Bhili Phonological Compatibility. To quantify segmental overlap, we computed phoneme inventory similarity between Hindi and Bhili using eSpeak-based grapheme-to-phoneme conversion. Comparing inventories extracted from our parallel corpus, Hindi covers 83.3% of the Bhili inventory with a Jaccard similarity of 0.80 (Table 3). This suggests low risk of severe segmental distortion under Hindi-based synthesis, while leaving suprasegmental mismatch (intonation, rhythm, phrasing) as the primary source of dialect divergence. Examples of sentences are mentioned in table 1 and linguistic correspondences are mentioned in table 2.

BhiliVaani-Bridge: Synthetic Speech Synthesis. BhiliVaani-Core is dialect-authentic but substantially smaller than what is typically required for stable, intelligible neural TTS. We therefore construct BhiliVaani-Bridge (Table 4, a controlled synthetic corpus generated with Hindi TTS to increase training scale, explicitly treated as a *bridge* rather than a substitute for community speech Sharma et al. (2025).

Selecting a Hindi TTS system. We conducted pilot experiments across multiple platforms that support Hindi speech synthesis, including Google Cloud Text-to-Speech[5], Microsoft Azure TTS[6], Bhashini[7], and IIIT-Hyderabad Indic TTS[8] . Across trials, Google Hindi TTS produced the most intelligible renditions of Bhili text. A reduced speaking rate of 0.8× further improved clarity. While outputs retained a Hindi accent at the suprasegmental level, native listeners consistently understood the content.

Generation and Filtering Pipeline. Synthetic corpus construction followed a controlled pipeline designed to yield training-friendly utterances while minimizing synthesis artifacts:

1. **Text selection:** Extract approximately 4,000 Bhili sentences from the Bhili-Hindi-English Parallel Corpus (BHEPC). This corpus provided grammatically well-formed Bhili text spanning diverse semantic domains.
2. **Length-Based Filtering:** To avoid truncation and maintain synthesis quality, we filtered to sentences $\leq$270 characters (pilot-determined upper bound for consistent Google TTS phonetics), yielding $\sim$3,000 sentences.
3. **Initial Synthesis:** We synthesized the $\sim$3,000 sentences with Google Cloud Text-to-Speech (Hindi) at 0.8× speed, producing $\sim$7.5 hours of audio, using a single voice (female "hi-IN-Wavenet-A") for acoustic consistency.
4. **Duration-Based Filtering:** To reduce alignment instability from long utterances, we removed samples >10 s, resulting in 2,500 utterances totaling 6.0 hours; this retains scale while improving per-utterance training suitability.

After filtering, BhiliVaani-Bridge contains 2,500 utterances totaling approximately 6.0 hours of synthetic speech. Consistent with controlled augmentation, we keep the synthetic corpus *single-voice* to avoid introducing additional speaker variability at the augmentation stage.

Table 4: BhiliVaani-Bridge Speech Corpus Statistics

Characteristic	**Value**
Utterances	2,500
Total duration	6.0 hours
Max utterance duration	10 s
Max text length	270 characters

Human Validation of Synthetic Intelligibility. Because the synthetic corpus is used to improve clarity in later staged-training experiments, we validated whether Google Hindi TTS outputs are intelligible to native Bhili listeners despite Hindi-accented suprasegmentals. We conducted a two-stage validation.

1. **Linguistic quality assessment:** Two Bhili-Hindi bilingual linguists evaluated 200 randomly sampled synthetic utterances for grammatical preservation, phonemic accuracy, and prosodic naturalness. All samples passed grammatical preservation and phonemic accuracy, confirming stable segmental rendering of Bhili text.
2. **Native-speaker intelligibility MOS :** Six native Bhili speakers from Jhabua (4 male, 2 female), ages 28-54, rated 50 utterances each on a 5-point Likert scale (1-5). The synthetic corpus achieved MOS-Intelligibility $\mathbf{4.86 \pm 0.08}$, indicating near-perfect comprehension despite non-native prosody. Evaluators consistently described the output as "Bhili read with a Hindi accent," motivating the staged training strategy developed in subsequent sections: use synthetic speech to sharpen intelligibility while using the original corpus to preserve dialect grounding.

[5] `https://docs.cloud.google.com/text-to-speech/docs`
[6] `https://learn.microsoft.com/azure/ai-services/speech-service/text-to-speech`
[7] `https://tts.bhashini.ai/openapi/ui/` (Bhashini Text-to-Speech model)
[8] `https://bhaasha.iiit.ac.in/indic-tts/demo/`

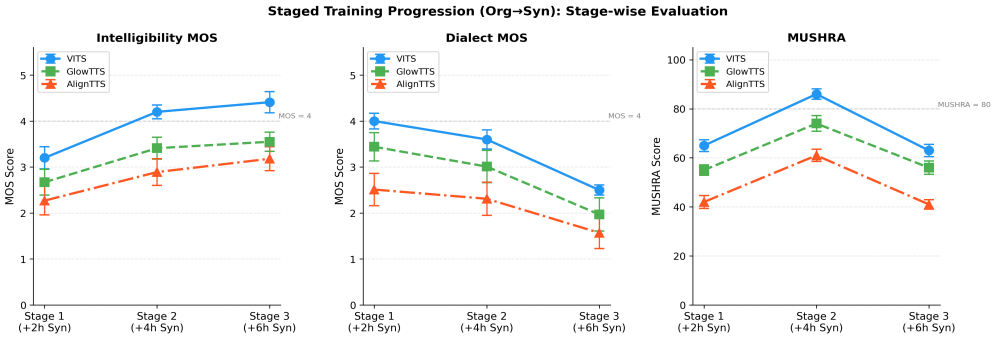

Figure 2: Stage-wise Intelligibility MOS, Dialect MOS, and MUSHRA scores for the proposed Original→Synthetic staged training framework across VITS, GlowTTS, and AlignTTS. Stage 1,2,& Stage 3 correspond to progressive fine-tuning with 2, 4, and 6 hrs of synthetic Bhili-proximate Hindi speech respectively. Error bars indicate standard deviation.

4 Evaluation Framework

Revitalization-oriented TTS requires more than intelligible speech: synthesized speech must also be recognized as locally authentic Bhili, otherwise it risks social misrecognition as "Hindi-like Bhili" and may be less acceptable for education, community media, and welfare-facing voice interfaces. We therefore evaluate generated speech along two perceptual axes: MOS-Intelligibility, which measures whether listeners can understand the utterance content, and MOS-Dialect, which measures whether the audio is perceived as locally Jhabua Bhili rather than Hindi-accented Bhili. Importantly, MOS-Dialect is judged from the synthesized audio itself; when intelligibility is low or prosody is unstable, listeners may down-rate dialect authenticity even if some Bhili phonetic cues remain. Thus, MOS-Dialect should be interpreted as perceived dialect authenticity under usable speech quality, not as isolated phonetic coloration. For MOS-Dialect, native Jhabua Bhili speakers who were also bilingual in Hindi were asked: "Does this audio sound like Jhabua Bhili, or like Hindi-accented Bhili?" using a 5-point scale, where 1 indicates Hindi-like speech and 5 indicates locally authentic Bhili speech. The same native-listener panel evaluated all systems under randomized blind presentation. Each configuration was evaluated on 50 held-out test utterances using MOS-Intelligibility, MOS-Dialect, and MUSHRA-style relative quality. MUSHRA ratings were collected on a 0–100 scale using a natural community recording as reference and a deliberately degraded sample as a low-quality anchor, allowing us to capture perceived naturalness beyond MOS scores. We report mean scores as shown in Table 5 with 95% non-parametric bootstrap confidence intervals over utterance-level ratings. Raters failing attention checks or pre-specified consistency checks were excluded before computing final scores.

4.1 Training Setup

All models were trained using the CoquiTTS framework. As the framework does not support out of the box devanagri script, all Bhili texts were transliterated to Roman using `indic_transliteration` tool. A custom cleaner was also implemented that normalizes whitespace, corrects invalid tokens, and enforces consistent formatting prior to model ingestion. Training was conducted on a single NVIDIA V6000 GPU (32 GB) for a maximum of 2,000 epochs; checkpoints were saved every 50 epochs and validation was run every 40 epochs. We used Adam with a decaying learning-rate schedule (Lambda scheduler; initial LR 1×10^{-5}) and selected final checkpoints by minimum validation loss. Convergence times were approximately 80h (VITS), 144h (GlowTTS), and 115h (AlignTTS). Communities engaged in revitalization efforts frequently highlight that modern speech synthesis pipelines require GPU access and specialized expertise, which can limit local control and iteration. Our staged-training design is partly motivated by this constraint: it aims to extract maximal utility from limited real Bhili speech data while using synthetic augmentation as regulated supervision, reducing the marginal cost of improving intelligibility without requiring large-scale studio collection. We focus on VITS/GlowTTS/AlignTTS as stable few-hour baselines; FastSpeech2/multilingual baselines require

Table 5: Intelligibility and Dialect MOS scores (mean ± CI) for all training configurations and TTS models. The Staged (Org→Syn) group contains three progressive fine-tuning stages, each adding 2 hours of BhiliVaani-Bridge. Best Intelligibility and Dialect scores per model are **bold**, and result with good dialect and intelligibility at a same time is blue.

Model	Metric	Baseline Configurations								Staged Training (Org → Syn)					
		Original		Synthetic		Mixed		Syn→Org		Stage 1		Stage 2		Stage 3	
		Mean	CI	Mean	CI	Mean	CI	Mean	CI	Mean	CI	Mean	CI	Mean	CI
VITS	Intelligibility	2.30	0.18	**4.62**	**0.20**	3.80	0.21	3.10	0.34	3.20	0.24	4.20	0.15	4.41	0.23
	Dialect	1.01	0.34	2.13	0.20	4.00	0.30	3.90	0.18	**4.00**	**0.17**	3.60	0.21	2.50	0.11
GlowTTS	Intelligibility	1.85	0.35	**3.69**	**0.29**	3.34	0.16	2.53	0.27	2.67	0.28	3.41	0.24	3.55	0.21
	Dialect	0.89	0.13	1.78	0.37	3.20	0.38	3.13	0.23	**3.44**	**0.31**	3.01	0.35	1.97	0.36
AlignTTS	Intelligibility	1.53	0.21	**3.22**	**0.34**	2.72	0.34	2.21	0.38	2.27	0.31	2.89	0.29	3.18	0.26
	Dialect	0.77	0.33	1.23	0.47	2.13	0.41	1.91	0.43	**2.51**	**0.35**	2.31	0.36	1.57	0.34

reliable duration supervision and robust Bhili front-end normalization, which are unavailable, and are left for future work Ren et al. (2019).

4.2 Training Regimes for Managing Dialect Drift

BhiliVaani provides two complementary resources that define our learning setting and enable controlled study of dialect drift. BhiliVaani-Core comprises 3 hours of dialect-authentic Jhabua Bhili speech, serving as an anchor for local phonology and prosody but offering limited scale when used as the sole supervision source. BhiliVaani-Bridge adds 6 hours of synthetic Bhili speech, increasing supervision volume and often improving clarity, but it also carries Hindi-accented suprasegmentals that can weaken perceived dialect authenticity. Together, Core and Bridge allow us to examine an intelligibility-dialect identity trade-off that is central for preservation- and community-facing TTS. In extremely low-resource TTS, gains from additional data depend not only on *how much* supervision is added, but also on its *source* and the *order* in which it is introduced. We therefore compare five training regimes that combine BhiliVaani-Core and BhiliVaani-Bridge: *Original-only*, *Synthetic-only*, *Mixed* (joint training on the union), *Synthetic→Original* (pre-train on synthetic then fine-tune on original), and *Original→Synthetic staged refinement*. In the proposed *Original→Synthetic* staged refinement, models are first trained on BhiliVaani-Core to establish dialect-specific acoustic priors, then progressively fine-tuned with synthetic data in three increments that add 2 hours per stage (Stage 1: +2h, Stage 2: +4h, Stage 3: +6h total). We benchmark VITS, GlowTTS, and AlignTTS across all regimes (Table 5, Figures 1-2); we additionally report VITS-based objective diagnostics (Table 6, Figure 3) to support analysis and deployment-oriented discussion.

5 Experimental Design & Results

We quantify the central trade-off in preservation-oriented Bhili TTS: generated speech must remain recognizable as *Jhabua Bhili* while achieving intelligible, stable synthesis under only a few hours of real bhili speech supervision. We benchmark three non-autoregressive architectures (VITS, GlowTTS, AlignTTS) across the five regimes in Section 4.2. Across architectures, the two corpora induce complementary behaviors: BhiliVaani-Core anchors local identity but is too small for consistently intelligible synthesis, whereas BhiliVaani-Bridge reliably improves intelligibility but imports Hindi-like suprasegmentals, reducing dialect recognizability. This motivates other training paradigms as a mechanism to improve clarity while controlling dialect drift.

1. **Original-Only Training (Dialect Anchor under Data Scarcity)**

   We train each model from scratch using only BhiliVaani-Core (3h). Original-only training preserves some Jhabua-specific acoustic coloration, but outputs are often difficult to follow due to unstable alignment, blurred segment boundaries, and inconsistent phrasing. Accordingly, intelligibility scores are low for all models (Table 5): VITS reaches 2.30±0.18, GlowTTS 1.85±0.35, and AlignTTS

Table 6: Objective evaluation metrics for VITS across training configurations. VAD-SNR measures signal quality on active speech segments; ΔC50 measures speech clarity relative to reference speech (closer to 0 is better); F0 Mean and Std characterise pitch distribution. "Reference" speech denotes test set data.

Samples from	**VAD-SNR (dB)**	**ΔC50 (dB)**	**F0 Mean (Hz)**	**F0 Std (Hz)**
Original	36.81	-7.99	165.76	35.56
Synthetic→Original	34.61	-4.26	161.58	32.41
Synthetic	33.53	-7.16	133.12	26.48
Mixed	36.38	-12.52	156.78	50.84
Stage 1	32.01	-3.06	129.39	21.39
Stage 2	33.36	-6.69	125.94	22.33
Stage 3	33.09	-6.23	142.78	27.41
Reference	30.13	0.00	168.46	36.45

1.53±0.21 in MOS-Intelligibility. Dialect scores are also low (VITS 1.01±0.34, GlowTTS 0.89±0.13, AlignTTS 0.77±0.33). This is expected under our evaluation definition: when speech is unstable and difficult to understand, dialect authenticity is down-rated even if some dialectal cues are present. This baseline shows that 3h of real bhili speech supervision is insufficient for community-usable intelligibility, motivating augmentation.

2. **Synthetic-Only Training (Intelligibility Gains and Dialect Drift)**

   We train using only BhiliVaani-Bridge (6h, 2,500 utterances). Synthetic-only training yields the strongest intelligibility across all architectures (Table 5): VITS achieves 4.62±0.20, GlowTTS 3.69±0.29, and AlignTTS 3.22±0.34. However, dialect authenticity drops substantially (VITS 2.13±0.20, GlowTTS 1.78±0.37, AlignTTS 1.23±0.47), consistent with synthetic speech carrying Hindi-like suprasegmental patterns. Thus, synthetic supervision resolves the alignment and clarity bottleneck but introduces undesirable dialect drift for preservation-oriented deployment.

3. **Mixed Training (Joint Optimization Effects)**

   We train on the combined data of BhiliVaani-Core and BhiliVaani-Bridge (9h total) by shuffling all utterances and optimizing jointly. Mixed training improves intelligibility over original-only for all models and yields strong dialect authenticity in the presence of real bhili speech. For VITS, Mixed achieves 3.80±0.21 (Intelligibility) and 4.00±0.30 (Dialect), which is the highest dialect MOS observed for VITS in Table 5 (tied with Stage 1). GlowTTS reaches 3.34±0.16 intelligibility and 3.20±0.38 dialect, and AlignTTS reaches 2.72±0.34 intelligibility and 2.13±0.41 dialect. Despite strong mean dialect scores, joint optimization can encourage distributional averaging; in experimentation this yielded utterance-level variability where synthetic prosody occasionally bleeds into phrasing. This motivates curricula that explicitly control exposure order.

4. **Synthetic→Original Fine-tuning (Partial Recovery after Synthetic Pretraining)**

   We evaluate whether dialect identity can be recovered after learning clarity from synthetic speech by pre-training on BhiliVaani-Bridge and fine-tuning on BhiliVaani-Core. This strategy partially restores dialect cues but does not fully eliminate synthetic priors. For VITS, Synthetic →Original yields 3.10±0.34 intelligibility and 3.90±0.18 dialect. For GlowTTS, the scores are 2.53±0.27 and 3.13±0.23; for AlignTTS, 2.21±0.38 and 1.91±0.43 (Table 5). The synthetic pretraining establishes strong timing and intonation priors that are difficult to fully overwrite with only 3h of real bhili speech, motivating the staged curriculum where dialect is anchored first.

5. **Original →Synthetic Staged Training (Curriculum Control of Dialect Drift)**

   Original→Synthetic staged training treats synthetic speech as a controlled refinement signal rather than a substitute for community recordings. Training proceeds in an initial dialect-grounding phase on BhiliVaani-Core, followed by three progressive fine-tuning stages that add synthetic speech in 2h increments (Stage 1: +2h, Stage 2: +4h, Stage 3: +6h).

Table 7: Manual error analysis results for VITS across training configurations. Failure rate (%) denotes utterances containing at least one major error. S: substitution, D: deletion, M: mispronunciation, U: unintelligible, R: repetition, DC: dialect collapse, F: failure rate.

Config.	**S**	**D**	**M**	**U**	**R**	**DC**	**F (%)**
Original	9	15	11	16	15	8	64
Synthetic	0	0	3	0	1	5	20
Mixed	2	1	2	1	1	4	16
Syn→Org	2	1	2	1	1	5	20
Stage 1	2	2	3	3	1	4	28
Stage 2	1	1	1	1	1	3	**12**
Stage 3	2	2	2	1	1	5	20

Stage-wise perceptual trajectory. Figure 2 shows a consistent pattern across models: intelligibility increases monotonically with added synthetic hours, while dialect authenticity peaks early and then declines as synthetic influence dominates suprasegmentals. For VITS, Stage 1 yields 3.20±0.24 intelligibility with 4.00±0.17 dialect; Stage 2 improves intelligibility to 4.20±0.15 while retaining moderate dialect authenticity at 3.60±0.21; Stage 3 further increases intelligibility to 4.41±0.23 but dialect drops sharply to 2.50±0.11 (Table 5). The same monotone intelligibility / non-monotone dialect trend is visible for GlowTTS and AlignTTS. For preservation-oriented deployment, we recommend Stage 2 as the default operating point because it provides a strong balance between usability and recognizability (for VITS:4.20±0.15 intelligibility with 3.60±0.21 dialect), while avoiding the pronounced dialect drift observed at Stage 3.

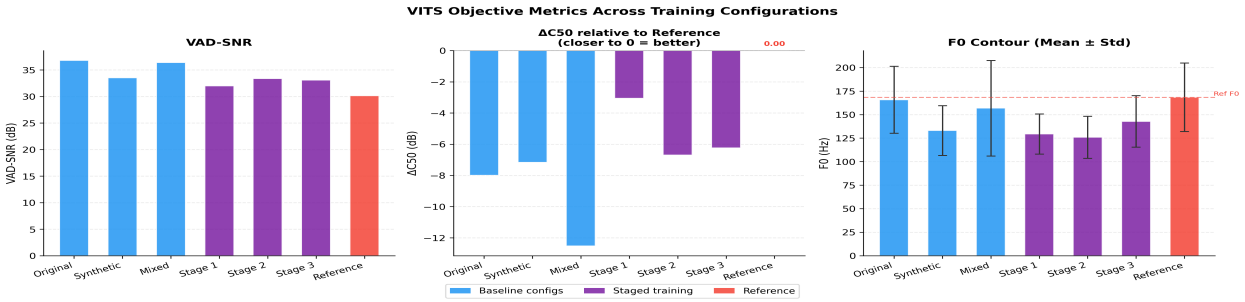

Figure 3: Objective metrics for VITS synthesized speech across training configurations. VAD-SNR measures signal quality on active speech segments; C50 measures speech clarity; F0 Mean ± Std characterizes pitch level and prosodic variation. Purple bars indicate the Original→Synthetic staged training configurations.

5.1 Objective Diagnostics for VITS

Objective diagnostics for VITS (Table 6) help interpret perceptual trends. VAD-SNR is higher for Original and Mixed (36.81 and 36.38 dB) than for staged configurations (32.01–33.36 dB), indicating that staging gains in intelligibility are not explained solely by cleaner signals. C50 becomes *less negative* in Stage 1 (−3.06) than in Original (−7.99) and Mixed (−12.52), consistent with improved perceived clarity when synthetic supervision is introduced. A key limitation of C50 is that it measures only the ratio of early to late sound energy and ignores frequency-dependent effects and background noise that strongly influence actual speech intelligibility. Therefore, the actual perceived speech of Mixed and Stage 2 trainings (Table 5 gives good results but contradicts with C50 measure. Prosodic statistics shift systematically with synthetic exposure: synthetic-heavy settings reduce mean F0 relative to Original (133.12 Hz vs. 165.76 Hz), and staged training moves toward the synthetic pitch distribution at Stage 1/2 (129.39/125.94 Hz), with partial recovery at Stage 3 (142.78 Hz), mirroring the observed dialect-prosody trade-off Mauch & Dixon (2014).

5.2 Error Analysis

Manual error analysis (table 7) was performed on 25 randomly sampled utterances per configuration, evaluated by two native Bhili speakers. Utterances were labeled for substitution, deletion, mispronunciation, unintelligible segments, repetition, and dialect collapse. An utterance was marked as failed if any major error occurred. Stage 2 achieves the lowest failure rate (12%), with only minor mispronunciations and occasional dialect collapse, indicating the best overall intelligibility and pronunciation. Mixed training follows (16%), benefiting from real Bhili exposure but still showing some dialect collapse. Synthetic-only and Synthetic→Original both record 20% failures: the former mainly reflects Hindi-style prosody, while the latter introduces additional substitutions during fine-tuning. Stage 3 (20%) and Stage 1 (28%) perform slightly worse, suggesting that both excessive synthetic exposure and insufficient supervision degrade quality. Original-only training performs the worst (64%), with frequent deletions and unintelligible speech due to the limited three-hour corpus.

5.3 Comparison Across Training Regimes

Table 5 and Figures 1-2 support three consistent conclusions across VITS, GlowTTS, and AlignTTS: (i) synthetic augmentation reliably improves intelligibility, (ii) unregulated synthetic exposure weakens dialect authenticity, and (iii) curriculum scheduling provides an explicit control knob that yields a tunable frontier between intelligibility and dialect recognizability. Among staged settings, Stage 2 provides the most practical preservation-oriented trade-off, supporting our hypothesis that, under extreme scarcity, *training order functions as a dialect-preservation mechanism.*

5.4 "Why Staged Training Reduces Dialect Drift"

Staged refinement succeeds because it separates *what the model should sound like* (dialect identity) from *how clearly it should speak* (alignment and articulation). Training first on BhiliVaani-Core establishes an acoustic prior for local segmental and prosodic cues. Subsequent synthetic exposure improves alignment stability and articulation, while sequential scheduling and reduced learning rates limit catastrophic overwriting of the dialect prior. In contrast, Mixed training optimizes conflicting distributions simultaneously and can converge toward acoustic averaging, while Synthetic →Original attempts to reshape a strong synthetic prosodic prior using only 3h of real bhili speech, which is often insufficient to consistently reorient timing and intonation.

6 Generalizability and Reproducibility

The results suggest a practical workflow for preservation-oriented TTS in extremely low-resource settings: (i) curate a small, community-governed dialect anchor corpus; (ii) bootstrap intelligibility using synthetic speech from a genetically related high-resource TTS; (iii) validate synthetic intelligibility with native listeners; and (iv) treat synthetic data as *regulated supervision* by controlling both *amount* and *timing* of exposure via staged curricula (with an intermediate stage such as Stage 2 serving as a default operating point). Importantly, evaluation should explicitly measure dialect authenticity alongside intelligibility, since systems can improve clarity while becoming less locally recognizable. This pipeline is reproducible and transferable to other endangered languages where a related high-resource neighbor exists.

7 Conclusion

We introduced BhiliVaani, a community-driven TTS corpus and benchmark for *Bhili* language, that operationalizes a preservation-critical requirement often overlooked in low-resource synthesis: generated speech should be both *understandable* and *locally recognizable* for revitalization-facing applications such as education, storytelling, accessibility, and public-information delivery. We constructed two complementary resources BhiliVaani-Core and BhiliVaani-Bridge and used them to study how intelligibility gains from cross-lingual bootstrapping interact with dialect drift. Across three non-autoregressive architectures (VITS, GlowTTS, AlignTTS) and five training regimes, we observed a consistent pattern: training on the small community corpus alone preserves dialectal evidence but yields unstable, weakly intelligible synthesis, whereas

synthetic-only scaling improves intelligibility while importing Hindi-like suprasegmentals that reduce perceived dialect authenticity. Excessive synthetic exposure can *hurt* real-test authenticity, with diminished intelligibility returns and increased dialect-collapse errors. We show that the *order and quantity* of synthetic supervision are practical control mechanisms. The proposed Original→Synthetic staged curriculum anchors a dialectal prior on real bhili speech and then introduces synthetic supervision in calibrated increments, producing a tunable frontier between intelligibility and authenticity rather than uncontrolled averaging. For preservation-oriented deployment, we recommend Stage 2 as the default operating point, as it delivers strong intelligibility while retaining substantially higher dialect authenticity than later stages. BhiliVaani will be released openly to support reproducible, community-centered research and to enable future extensions to broader speakers, domains, and dialects.

8 Limitations

BhiliVaani reflects preservation-oriented TTS under extreme scarcity. BhiliVaani-Bridge is produced by a Hindi TTS system: it improves segmental clarity but imports Hindi suprasegmentals, so dialect drift at higher synthetic exposure is expected and motivates our Stage 2 recommendation; improving beyond this likely requires dialect-aware constraints (e.g., prosody normalization/regularization). BhiliVaani-Core is small and Jhabua-specific; broader speakers, spontaneous speech, and culturally salient genres are needed for stronger generalization. Responsible use requires consent, fair compensation, transparency about synthetic augmentation, and iterative community feedback.

9 Ethics and Broader Impact

This work involves Bhili speech recordings and native-speaker evaluations. All speakers, annotators, and listening-test participants provided informed consent for their respective roles and were compensated. Personally identifying metadata was removed or pseudonymised, and any public release of data or models will follow participant consent and applicable licensing constraints. Since speech can reveal speaker identity and regional background, released materials avoid speaker-identifying information and include clear documentation of intended use. BhiliVaani-Bridge contains synthetic speech generated by applying Hindi TTS to Bhili text. These utterances are used only as augmentation data and are explicitly labeled as synthetic. We do not present them as natural Bhili community recordings. The resource is focused on the Jhabua dialect and should not be interpreted as representing all Bhili varieties. Main risks include overgeneralising one dialect and mislabeling Hindi-accented synthetic speech as authentic Bhili; we mitigate these through provenance labels, documentation, and cautious release terms.

References

Cristian Ahumada, Claudio Gutierrez, and Antonios Anastasopoulos. Educational tools for mapuzugun. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pp. 183–196, 2022.

Srija Anand, Praveen Srinivasa Varadhan, Mehak Singal, and Mitesh M Khapra. Elaichi: Enhancing low-resource tts by addressing infrequent and low-frequency character bigrams. *arXiv preprint arXiv:2410.17901*, 2024.

Rohan Badlani, Adrian Łańcucki, Kevin J Shih, Rafael Valle, Wei Ping, and Bryan Catanzaro. One tts alignment to rule them all. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6092–6096. IEEE, 2022.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33: 12449–12460, 2020.

Shashi Bhushan, Ved Prakash Mishra, Vinay Rishiwal, Sharmila Arunkumar, and Udit Agarwal. Advancing text-to-speech systems for low-resource languages: Challenges, innovations, and future directions. *IEEE Access*, 2025.

Steven Bird. Decolonising speech and language technology. In *Proceedings of the 28th international conference on computational linguistics*, pp. 3504–3519, 2020.

Marcely Zanon Boito, Antonios Anastasopoulos, Marika Lekakou, Aline Villavicencio, and Laurent Besacier. A small griko-italian speech translation corpus. *arXiv preprint arXiv:1807.10740*, 2018.

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International conference on machine learning*, pp. 2709–2720. PMLR, 2022.

Edresson Casanova, Sandra Aluísio, and Moacir Antonelli Ponti. Tts applied to the generation of datasets for automatic speech recognition. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 1*, pp. 633–638, 2024.

Yuan-Jui Chen, Tao Tu, Cheng-chieh Yeh, and Hung-Yi Lee. End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning. In *Proc. Interspeech 2019*, pp. 2075–2079, 2019.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 6022–6034, 2020.

Isaac Elias, Heiga Zen, Jonathan Shen, Yu Zhang, Ye Jia, RJ Skerry-Ryan, and Yonghui Wu. Parallel tacotron 2: A non-autoregressive neural tts model with differentiable duration modeling. *Interspeech 2021*, 2021.

Tūreiti Keith. Work in progress: Text-to-speech on edge devices for te reo māori and ʻōlelo hawai i. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages@ LREC-COLING 2024*, pp. 421–426, 2024.

Youngjae Kim, Yejin Jeon, and Gary Lee. Audio-based linguistic feature extraction for enhancing multilingual and low-resource text-to-speech. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13984–13989, 2024.

Gokul Karthik Kumar, SV Praveen, Pratyush Kumar, Mitesh M Khapra, and Karthik Nandakumar. Towards building text-to-speech systems for the next billion users. In *Icassp 2023-2023 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 1–5. IEEE, 2023.

Florian Lux, Julia Koch, and Ngoc Thang Vu. Low-resource multilingual and zero-shot multispeaker tts. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 741–751, 2022.

Matthias Mauch and Simon Dixon. pyin: A fundamental frequency estimator using probabilistic threshold distributions. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 659–663. IEEE, 2014.

Aidan Pine, Dan Wells, Nathan Brinklow, Patrick Littell, and Korin Richmond. Requirements and motivations of low-resource speech synthesis for language revitalization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7346–7359, 2022.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52, 2024.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, 32, 2019.

Sebastian Ruder, Anders Søgaard, and Ivan Vulić. Unsupervised cross-lingual representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pp. 31–38, 2019.

Ashwin Sankar, Srija Anand, Praveen S Varadhan, Sherry Thomas, Mehak Singal, Shridhar Kumar, Deovrat Mehendale, Aditi Krishana, Giri Raju, and Mitesh Khapra. Indicvoices-r: Unlocking a massive multilingual multi-speaker speech corpus for scaling indian tts. *Advances in Neural Information Processing Systems*, 37:68161–68182, 2024.

Ashwin Sankar, Yoach Lacombe, Sherry Thomas, Praveen Srinivasa Varadhan, Sanchit Gandhi, and Mitesh M Khapra. Rasmalai: Resources for adaptive speech modeling in indian languages with accents and intonations. *arXiv preprint arXiv:2505.18609*, 2025.

Divya V Sharma, Vijval Ekbote, and Anubha Gupta. Indicsynth: A large-scale multilingual synthetic speech dataset for low-resource indian languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22037–22060, 2025.

Abhayjeet Singh, Amala Nagireddi, Jesuraja Bandekar, Sandhya Badiger, Sathvik Udupa, Prasanta Kumar Ghosh, Hema A Murthy, Pranaw Kumar, Keiichi Tokuda, Mark Hasegawa-Johnson, et al. Limmits'24: Multi-speaker, multi-lingual indic tts with voice cloning. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pp. 61–62. IEEE, 2024.

Pooja Singh, Sandeep Chatterjee, Gullal S Cheema, Amrit Singh Bedi, Tanmoy Chakraborty, Sandeep Kumar, and Ankita Shukla. Findings of the mmloso 2025 shared task on machine translation into tribal languages. In *Proceedings of the 1st Workshop on Multimodal Models for Low-Resource Contexts and Social Impact (MMLoSo 2025)*, pp. 121–129, 2025.

Praveen Srinivasa Varadhan, Ashwin Sankar, Giri Raju, and Mitesh M Khapra. Rasa: Building expressive speech synthesis systems for indian languages in low-resource settings. In *Proc. Interspeech 2024*, pp. 1830–1834, 2024.

Lucas H Ueda, Leonardo Marques, Flávio Simões, Mário Uliani Neto, Fernando Runstein, Bianca Dal Bó, and Paula DP Costa. Exploring synthetic data for cross-speaker style transfer in style representation based tts. In *Proc. SynData4GenAI 2024*, pp. 76–80, 2024.

Suhua Wang, Zifan Wang, Xiaoxin Sun, DJ Wang, Zhanbo Liu, and Xin Li. Manchutts: Towards high-quality manchu speech synthesis via flow matching and hierarchical text representation. *arXiv preprint arXiv:2512.22491*, 2025.

Xueyao Zhang, Yuancheng Wang, Chaoren Wang, Ziniu Li, Zhuo Chen, and Zhizheng Wu. Advancing zero-shot text-to-speech intelligibility across diverse domains via preference alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12251–12270, 2025.